UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Developing Methods for Construction of Population Pedigrees From Low Coverage Sequencing Data

A thesis submitted in partial satisfaction of
the requirements for the degree of

MASTER OF SCIENCE

in

BIOMOLECULAR ENGINEERING &

BIOINFORMATICS

by

Erik Hanson

March 2020

The Thesis of Erik Hanson is
approved:

_____
Professor Russell Corbett-Detig, Chair

_____
Professor Angela Brooks

_____
Professor Christopher Vollmers

_____
Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures & Tables

Abstract

Developing Methods for Construction of Population Pedigrees From Low Coverage Sequencing Data

by **Erik Hanson**

A population pedigree is a graph that captures the totality of the family and genetic histories within a population. While pedigrees contain an abundance of advantageous information for genomic studies, assembling one is often tedious, time consuming, and fraught with error. A combination of highly multiplexed low-coverage sequencing, genotype imputation, and relationship inference software makes it feasible to develop a pedigree cheaply and efficiently. By applying this approach to an experimental admixed *Drosophila melanogaster* population we developed a dataset that contains genome-wide variants for thousands of individuals in our population. We were also able to confidently identify over one thousand parent-offspring relationships from almost four thousand sequenced samples. However, we were not able to construct a complete pedigree due to overestimates of relatedness resulting from our population's mixed ancestry. Implementing software that account for population structure could rectify this issue and provide more accurate relationship inference within our population.

Acknowledgements

Foremost, I would like to pay special regards to my committee chair and advisor, Dr. Russell Corbett-Detig. His expertise and guidance were vital assets to the completion of this project. His enthusiasm and excitement for the project continuously motivated me throughout the processes of data generation, analysis, and the writing of this thesis. I could not have imagined a better advisor for this project.

Additionally, I would like to thank the other members of my thesis committee: Dr. Angela Brooks and Dr. Christopher Vollmers, for their feedback and insightful comments.

My completion of this project could not have been accomplished without the assistance of my colleagues: Max Genetti and Evan Pepper. Their aid in preparation of libraries was invaluable to generating our data set consisting of thousands of individuals. Additionally, Max provided meaningful contributions towards processing preliminary sequencing results and alignments.

My sincere thanks goes to Niki Thomas, for assistance with editing, formatting, and figure generation. Her contributions were critical in adding clarity and cohesion to the writing of this thesis.

I am extremely grateful for the support provided by my parents, Mark and Linda Hanson. Their limitless love and encouragement were essential for the completion of my Master's Degree.

And finally, I would like to give special thanks to every member of the Corbett-Detig lab. Not Trash!

Background

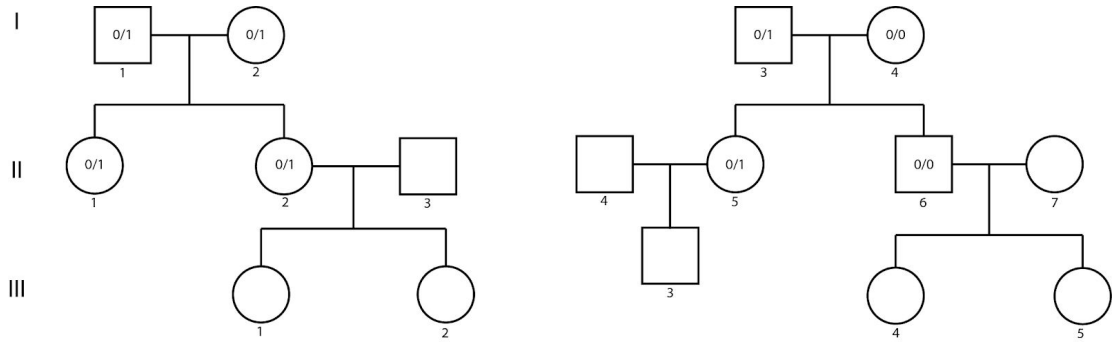

Figure 1: A pedigree graph depicting two unrelated families within the same population. Square nodes represent males, while circle nodes represent females. A subset of individuals contain genotypes which are located at the same autosomal bi-allelic variant locus.

A population pedigree is a graph that represents chronological familial relationships between members of a population (**Fig. 1**). The dynamic history of genetic lineages can be traced through the structure of a pedigree, which enables analyses such as genome wide association tests [Lu et al., 2016; Tore et al., 2011], shifting genetic variation [Chen et al., 2019], and heritable mutations/polymorphisms [Okinaka et al., 1997; Ouweland et al., 1992]. Therefore, a complete population pedigree is a rich and powerful data set which wholy encapsulates the genetic record of a population and its individuals. While pedigrees provide a plethora of useful information, they are traditionally developed by manually recording familial histories. This process is often laborious and error prone [Kerr et al., 2013; Lathrop et al., 1983], making it especially impractical for use on large genomic data sets. Combined with a large multi-generational population the task of constructing a pedigree quickly becomes infeasible. However, a combination of next generation sequencing and

modern computational techniques make it possible to infer relationships and construct pedigrees from large scale SNP genotyping data.

Table 1: Expected values of the proportion of the genome that is identical by descent (IBD) 0,1, and 2 in first and second degree familial relationships. IBD 0,1, and 2 refer to two individuals sharing 0,1,or 2 alleles that are identical by state and inherited from the same common ancestor. For example, for any given bi-alllelic loci, a parent and offspring share at least one allele by state and by descent because the child inherits the allele from their parent. In turn the parent inherited the same allele from the child's grandparent.

Relationship	Prop. IBD 0	Prop. IBD 1	Prop. IBD 2
Parent-Offspring	0	1	0
Full-Siblings	0.25	0.5	0.25
Half-Siblings	0.5	0.5	0
Grandparent-grandchild	0.5	0.5	0
Avuncular	0.5	0.5	0
First Cousins	0.75	0.25	0

These algorithms identify regions of the genome that are identical by descent to determine the level of relatedness between two individuals [Browning & Browning, 2013; Epstein et al., 2000; Purcell et al., 2007; Schaffner et al., 2018]. Given a familial relationship, there is an expected proportion of the genome, between two individuals, that is shared by a common ancestor (**Table 1**). Using the states at bi-allelic sites, algorithms can calculate this proportion of identity by descent (IBD) and then compare it to the expected values to predict the relationship between two individuals. Additionally, these programs are often robust to both a large population size and number of variant sites.

Although this addresses one need in constructing large pedigrees, generating the input data set can still be a costly and time consuming endeavor. Reliably calling

genome wide variants and genotypes in an individual often requires high coverage sequencing [Nielsen et al., 2011; Song & Zhang, 2016]. Given a large population, the cost to deeply sequence each individual rapidly exceeds the budget of most NGS studies. One way to skirt this obstacle is to lightly sequence many individuals from a population. This can be achieved, *e.g.*, by utilizing multiplexing methods that allow for large sample pools to be sequenced at one time [Roland & Reich, 2012]. In conjunction with a Tn5 transposase library preparation that makes use of homemade enzymes [Picelli et al., 2014], the cost to sequence thousands of individuals lightly becomes quite affordable. This method of preparing and sequencing libraries minimizes hands-on time while simultaneously keeping the price per library cost-effective for large genomics studies [Hennig et al., 2018]. Additionally, the Tn5 enzyme can readily be applied to a broad range of applications since the transposase can anneal to user designed oligonucleotides [Picelli et al., 2014].

Algorithms designed to impute genotypes from low coverage data can address the lack of genome wide coverage procured by this approach. Imputation algorithms generally operate under the assumption that an individuals' genome is made up of a mosaic of the founding haplotypes within the population. By comparing reference haplotypes to an individual's sequence data, they determine windows of the genome that represent a reference haplotype. The allelic states of the haplotype along this window can then be used to impute missing genotypes. These algorithms can be applied by using previously developed reference haplotypes [Browning & Browning,

2016; Howie et al., 2012; Li et al., 2010; O'Connell et al., 2014] or by reconstructing haplotypes from population data, in the absence of a reference panel. STITCH is one such method that accomplishes imputation through the reconstruction of haplotypes from low coverage data [Davies et al., 2016]. STITCH has been shown to accurately impute genotypes with coverages < 1x [Davies et al., 2016; Zan et al., 2019], so long as the constituent haplotypes underlying these data are observed a sufficient number of times within the sampled population. Combining low coverage sequencing with subsequent genotype imputation can affordably provide a great deal of information to be used for pedigree construction and downstream analysis.

The population of focus in this study is an experimental, admixed *Drosophila melanogaster* population generated in our lab. The two related populations, originating from France [Pool et al., 2012] and Ethiopia [Lack et al., 2015], are recently isolated (~10,000 years). Furthermore, studies have demonstrated various forms of reproductive isolating incompatibilities between parapatric African and Cosmopolitan *Drosophila melanogaster* lineages [Hollocher et al., 1997; Lachance and True, 2010; Ting et al., 2001; Yukilevich & True, 2008]. The recent divergence and isolating factors between the two populations make this hybrid population ideal for studying incipient speciation. Additionally, we can use this population to investigate the genomic effects of admixture between African and Cosmopolitan *Drosophila* species. This would further draw insight on the origins of modern North American *Drosophila melanogaster* species, which have been shown to possess

composite genomes developed by admixture between African and European populations [Pool, 2015].

In this study, we use cost-effective methods to sequence several thousand members of our population and work towards developing a whole population pedigree. We applied the Tn5 library prep method, followed by genotype imputation, to cheaply and efficiently procure a genome wide dataset representative of several thousand individuals within our population. Using this information, we made use of pairwise relationship inference programs to determine the relationships between samples using solely genomic data. Finally, we investigate the validity of parent-offspring (PO) relationships detected in our population to determine the feasibility of constructing a population pedigree through our approach.

Methods

Fly Population

For each of the two founding populations, ten lines were kept in separate containers and allowed to propagate for 5 generations. This allowed for adequate recombination and mixing of the founding haplotypes within each respective population. Next, the populations were placed in the same container to form the F0 generation of our hybrid population. After the flies mated and produced eggs, the adult flies were removed and frozen at -20℃ for future DNA extraction. The eggs were then allowed to mature into adult flies ready to produce another generation. This greatly reduces the complexity of developing a population pedigree because it

guarantees that our generations have no overlap. Thus, reducing the dimensionality when trying to infer unknown relationships between members of the population. This process was repeated 7 more times until the population reached the F7 generation. Flies were fed and maintained under standard *Drosophila* practices.

Genomic DNA Extraction and Purification

Each fly's genomic DNA (gDNA) was extracted using a squish prep method. Each fly was individually submerged in 50µl of lysis buffer and 1µl of 500 mAU/mL Protease K to degrade cellular proteins. The flies were then crushed and lysed using a pipette tip. Next, the lysate for each fly was incubated at room temperature overnight, followed by 10 min at 95℃ to denature excess Protease K. After this, the lysate was subjected to a SPRI bead clean up, in which 90µl of SPRI beads were added to the lysate. Following a 5 minute incubation at room temperature, the lysate was removed and two rounds of >80% ethanol washes were performed. After drying the excess ethanol, the gDNA is eluted off the SPRI beads with 30 ul of ddH_2O.

Tn5 Library Preparation

In order to prepare libraries for short read paired-end Illumina sequencing we used a method adapted from Picelli et al. In separate tubes, the Tn5ME-A (5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3') and the Tn5ME-B (5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3') adapters are annealed to equal parts of the Tn5ME-R oligo (5'-[phos]CTGTCTCTTATACACATCT-3') at 95 ℃ for 5 minutes. Next, Tn5

transposase is added (1µl for every .143 µl of oligos) to each tube containing the annealed oligos. Incubation of the Tn5-oligo mixture occurs at room temperature for 1 hour to form the transposome complex. Next, genomic DNA (gDNA) is tagmented by adding 2µl of the mixed transposome complex, 4 µl of 5x TAPS-PEG buffer, 12 µl of H2O, and 2 µl of gDNA, for a total reaction volume of 20 µl. Following a 10 minute incubation at 55 ℃, 5 µl of 0.2% SDS is added to the reaction to denature the Tn5 enzyme, preventing further tagmentation of gDNA. Five microliters of the tagmentation reaction, containing the tagmented gDNA, is added to a KAPA HiFi (Cat. No. KK2102) PCR reaction mixture containing: 11.75µl of H2O, 5µl of 5x KAPA HiFi buffer, 0.75µl of dNTPs, 0.5µl of HiFi polymerase, and 2µl of a uniquely dual indexed i5 and i7 primer pair, for a total reaction volume of 25µl. This reaction is run at: 1 .72℃ for 5 min, 2. 95℃ for 3 min, 3. 98℃ for 20 sec, 4. 65℃ for 15 sec, 5. 72℃ for 30 sec, 6. Repeat steps 3 to 5 11x, 7. 72℃ for 5 min for a final extension phase.

Pools of 96 samples were created by mixing 1µl of each individual's amplified library. This pool was then cleaned and size selected for fragments >300bp using a Zymo Research Select-A-Size DNA Clean & Concentrator kit (Cat. No. D4080). The size-selected pools were then quantified using a Qubit 3 Fluorometer (dsDNA HS protocol) and an Agilent Tapestation 2000. Equi-molar amounts of 4 size-selected pools were mixed together to produce a "master pool" of 384 uniquely

dual indexed libraries. Each "master pool" was sent to Fulgent Genetics to be sequenced on a full lane of an Illumina Hi-Seq 4000 flow cell.

Sequencing QC

Raw paired-end reads for each sample were aligned to a chimeric genome consisting of the Release 6 plus ISO1 MT assembly of the *Drosophila* reference genome [dos Santos et al., 2015] and the GCF_000008025.1 assembly of the Wolbachia genome [Wu et al., 2004]. Alignments were generated with the BWA mem algorithm using default settings [Li, 2013]. Genome coverage was calculated by generating a pileup file using samtools mpileup [Li et al., 2009] with a minimum read and base quality of 30 (phred-scaled). Using the pileup, read depth was recorded at every 5000th basepair along each chromosome. The read depths were summed, multiplied by 150 (the length of the paired-end reads), and divided by the total genome length to calculate coverage. Coverage for an individual chromosome was calculated with the same methodology, except the genome length was substituted for the length of the chromosome.

The sex for an individual fly was determined by calculating the ratio between autsomal and X chromsome coverage. Since *Drosophila*'s genome is diploid, the heterogametic sex (males) only possesses one copy of the X chromosome, while the other sex will possess two copies. Therefore, the ratio of X chromosome coverage to diploid genome coverage should fluctuate based on the sex of an individual. Expected values for this ratio are: 0.5 for males and 1.0 for females.

Variant Calling and Imputation

ANGSD [Korneliussen et al., 2014] was used to call genome wide variants in our population using the parameters "-GL 2", "-doGlf 2", "-doMajorMinor 1", " -doMaf 2", and "-minMaf 0.01". Only variant sites with bi-allelic SNPs and a minor allele frequency (MAF) > 0.05 were used for subsequent analysis.

STITCH [Davies et al., 2016] was used to impute missing genotypes at variant sites called by ANGSD on each chromosome of the *Drosophila melanogaster* genome for individuals in our population. Each individual's alignment file and the bi-allelic variants located on a given chromosome were input into STITCH. STITCH was run with the parameters: founding haplotypes = 30 and 5 generations since the population has been founded. STITCH ran into memory issues when running 3754 samples across chromosomes larger than 10 Mb. Due to this, imputations on the major autosomal arms and the X chromosome were performed by chunking the chromosome into 10 Mb segments with 1Mb overlapping windows (ex. 1bp-10Mb,9Mb-19Mb,18-28Mb). Imputations along chromosomes smaller than 10 Mb were completed using one continuous segment spanning the entire chromosome. STITCH returns imputed genotypes for each individual in a multi-sample variant call format (VCF) file.

Each chromosome required to be split into overlapping segments was collapsed into one continuous VCF by masking genotypes that were not concordant between the two overlapping windows for a given individual. In other words, if an

individual had conflicting genotype imputations at a particular site within an overlapped region, then an individuals' genotype at that site was converted to a no call (ex. ./.). Imputed genotypes on the major autosomal arms and the X chromosome were filtered for an Info Score > 0.95 (reported by STITCH) and a minor allele frequency > 0.05. Info score is a common metric used to assess the quality of imputation that measures genotype information content. It ranges from 0 to 1, with a value 0.95 meaning that the imputed genotypes are equal to a dataset in which 95% of the samples have confidently identified genotypes. For each male individual, all heterozygous genotypes on the X chromosome were masked and removed from analysis. For each female, all genotypes on the Y chromosome were masked and removed from analysis.

The top 5 samples with the highest genome coverage and X chromosome coverage were utilized to benchmark STITCH's imputation accuracy within our population. All but one individual overlapped between the two groups. The samples' genotypes were called using bcftools mpileup and call with the parameter: minimum mapping quality of 20 [Li, 2011]. The genotypes called for each individual were compared to the genotypes imputed by STITCH on chromosomes: 2L, 2R, 3L, 3R, and X. Only common genome positions between the calls from STITCH and mpileup, where a genotype was successfully called from both outputs, were used for the comparison.

Pairwise Relationship Inference

In order to construct the pedigree, we set out to identify parent-offspring (PO) relationships within our population. This is due to the fact that the PO relationship has a clear and unique proportion of the genome shared by descent when compared to other familial relationships (**Table 1**). Furthermore, with our experimental population we were able to prevent generations from overlapping. This allows us to rule out any erroneous PO pairs identified within a generation (ex. F0-F0) or across two or more generations (ex. F0-F2).

PLINK [Purcell et al., 2007] was used to filter the genome wide variant sites output by STITCH and generate a bed file for input into relationship inference software. The parameters MAF > 0.1 and geno = 0.1 were used to filter variants for input. This geno parameter ensures that for each variant site output, at least 90% of individuals have a genotype called at that site. Using the same filters as above, PLINK's "--sample-diff counts-only" package performed pairwise calculations of the number of sites that have discordant homozygous genotypes between each individual in our population. This is an informative criterion for validating predicted PO relationships because, for any PO duo, each site compared should have at least one allele that is identical by state (IBS). An abundance of sites that are IBS 0 indicates that these two individuals do not likely represent a true PO relationship.

Pairwise relationship inference was performed with KING's related and ibdseg packages [Manichaikul et al., 2010]. This program takes a bed file produced by PLINK and attempts to determine the relationship between each pair of individuals

within the population. It compares the allelic states at each variant site and calculates the probability of sharing zero IBD and a kinship coefficient to predict the relationship between two population members. This robust software's ability to generate millions of relationship inferences in minutes makes it an viable option for identifying relationships with our large SNP data set and population.

Results

Sequencing QC

We prepared and sequenced the libraries of 3835 individuals within our population. These individuals span four generations (F0-F3). Out of the individuals sequenced, 3817 (99.5%) successfully returned raw reads. Only 3754 (97.9%) individuals had a genome coverage greater than 0.05x. These 3754 individuals make up our population for further downstream analysis. The population had an average genome coverage of ~1.2x, with the highest mean coverage being 6.3x (**Fig. 2**). Using the ratio of X coverage to autosome coverage, we identified 1872 males and 1882 females in our population of 3754 flies. The males had an average X:Autosome depth ratio of 0.51, with a range of 0.24. While females had an average ratio of 1.00, with a range of 0.53 (**Fig. 3**).

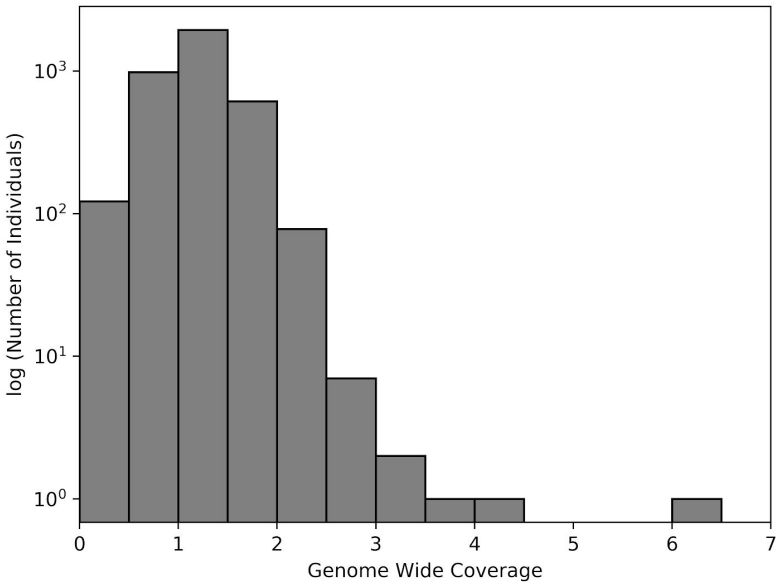

Figure 2: The distribution of genome coverages of all 3754 individuals in our population.

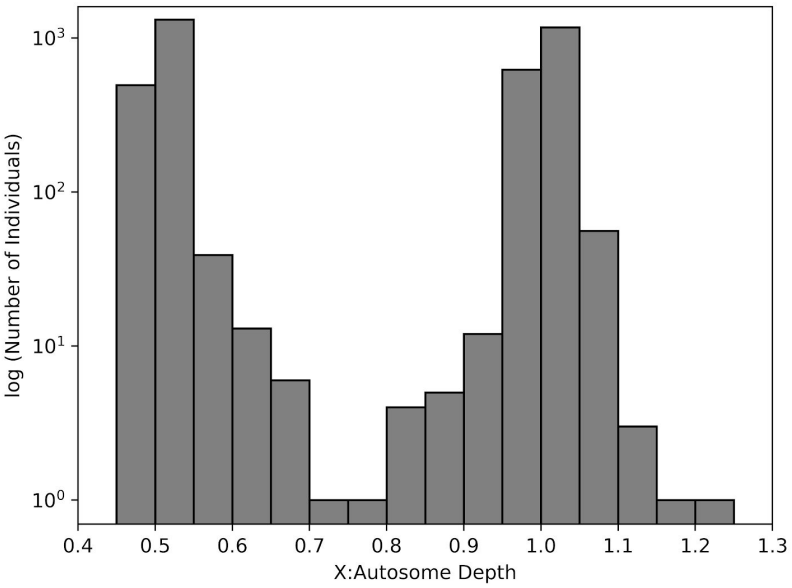

Figure 3: The distribution of X to Autosome depth ratio in our population. This ratio was used to determine the sex of the members of the population. A clear bimodal distribution formed between the male and female flies, with males having an expected value of 0.5 and females 1.0.

Variant Calling and Imputation

In total, ANGSD detected variants at 2,333,908 sites spread across the whole genome with a MAF > 0.05. After filtering for info score, STITCH was able to confidently impute genotypes at 1,884,550 of the variant sites called by ANGSD. Imputed genotypes at these sites were used for downstream pairwise relationship inference analyses.

Table 2: The results of the concordance comparison between STITCH and bcftools on the 4 major autosomal arms of the *Drosophila* genome. Discordant sites are sites where STITCH and bcftools produced non agreeing genotypes (ex. 0/0 and 0/1). Hom-Het represents a site in which bcftools genotype calling resulted in a homozygous genotype and STITCH's imputation resulted in a heterozygous genotype. % Hom-Het refers to the percentage of discordant sites that were Hom-Het.

Sample	% Discordant	% Hom - Het	Total Sites	Coverage
F0-P7-D12	13.1	73.2	1167288	6.2
F0-P8-H7	4.7	86.9	1168383	4
F0-P9-D10	13.8	77.5	1167736	3.5
F1-P11-D9	7.2	68.8	1168065	3.2
F0-P7-D10	12.2	77.4	1167389	3

We used bcftools to call genotypes for the 5 individuals with the highest genome-wide and X-chromosome coverage to assess STITCH's imputation accuracy. These individuals had genome coverages ranging from 6.2x to 3x; and X chromosome coverages ranging from 6.2x to 1.6x (**Table 2**, **Table 3**). This resulted in 1,951,683 loci with genotypes called on the X chromosome and the 4 major autosomal arms. The 5 individuals with the highest genome coverage had, on average 1,167,772 sites across the major autosomal arms in which STITCH and bcftools

successfully produced a genotype. On average, ~10.2% of compared sites on the autosomal arms were discordant between genotypes called by STITCH and bcftools (**Table 2**). Similarly, for the 5 individuals with the highest X coverage, on average ~11% of sites of compared sites on the X chromosome were discordant.

Table 3: The results of the concordance comparison between STITCH and bcftools on the X chromosome. The 5 highest coverage individuals had the highest coverages on the X chromosome within the population. The lowest coverage individual is a high genome coverage male. Discordant sites are sites where STITCH and bcftools produced non agreeing genotypes (ex. 0/0 and 0/1). Het-Hom is a site in which bcftools produced a heterozygous genotype, while STITCH produced a homozygous genotype. % Hom-Het and %Het-Hom refers to the percentage of discordant sites that were Hom-Het or Het-Hom.

Samples	% Discordant	% Hom-Het	% Het-Hom	Total Sites	X Coverage
F0-P7-D10	11.9	42.5	51.0	220126	6.2
F0-P7-D12	6.0	90.0	8.5	220016	4
F0-P8-A10	16.0	24.3	68.5	220225	3.5
F0-P8-H7	10.5	87.1	11.5	220084	3.2
F0-P9-D10	12.9	68.5	27.5	219211	3
F1-P11-D9	29.2	0	85.9	220307	1.6

For the individuals with the highest genome coverage, the majority of discordant sites (mean of 76%) were the result of homozygous calls by bcftools and heterozygous imputations by STITCH (**Table 2**). Only three individuals used for the X chromosome comparison followed the same pattern. On the X chromosome, the only male individual (F1-P11-D9) included in this comparison contained no discordant sites following this pattern (**Table 3**). Instead the large majority of this individual's discordant sites were the result of heterozygous calls by bcftools and homozygous imputations by STITCH.

Pairwise Relationship Inferences

After filtering with PLINK, a total of 1,133,516 variant sites were kept for relationship inferences. King predicted a total of 6885 PO duos within our population. Out of these PO pairs, 2056 were identified within the same generation. Given the discrete nature of the generations in our population, these relationships were not considered for further analyses. Out of the remaining 4829 PO relationships: 1687 resulted in an offspring being assigned more than two parents; 1156 assigned an offspring two parents of the same sex; 1201 only found one parent for an offspring; and 784 of them yielded a correct PO trio in which the parents consisted of one male and one female. Only the one parent PO duos and correct PO trios were considered for further analysis because they were the only sets of biologically possible relationships.

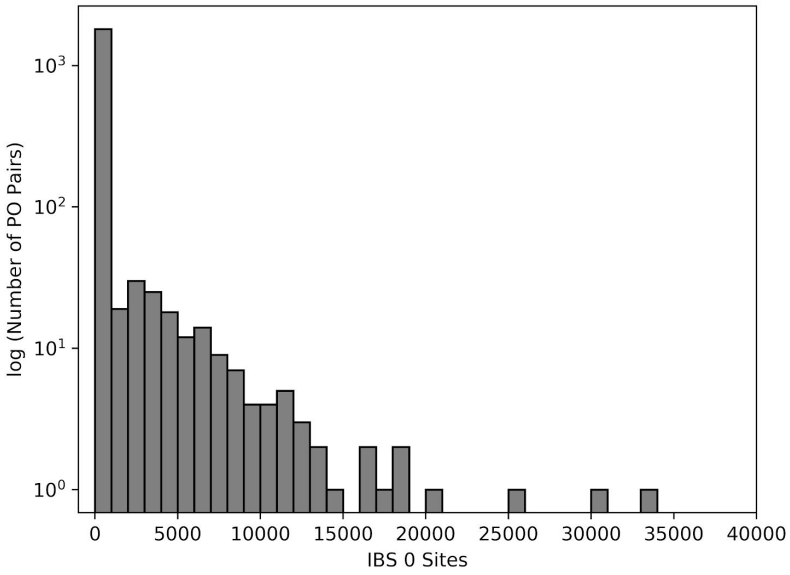

Figure 4: The distribution of the amount of IBS 0 sites for the 1979 biologically valid PO pairs analyzed. PO pairs should share half of the whole genome by descent, so they should always have one allele in common at each site compared (excluding sequencing and imputation error). This means that they should not have any sites that are IBS 0.

In order to determine the validity of the 1985 biologically valid assigned PO relationships, I examined the number of sites that were IBS 0 between each PO pair. For any PO duo, we would expect to see 0 sites that are IBS 0 between the parent and offspring. Only 93 PO duos met this expectation. However, ~92% of PO pairs had less than 1000 sites that were IBS 0 (**Fig. 4**). This is within a range of genotyping and sequencing error, given an average of 1,131,173 sites compared between each duo. The highest number of IBS 0 sites between predicted PO relationships was 33199, suggesting that this is likely a false positive PO relationship.

Additionally, I compared the number of IBS 0 sites between each offspring from a PO duo and each individual in the parental generation that was not the offspring's assigned parents. I would expect this comparison to yield an abundance of sites that are IBS 0 across the genome. Out of the approximately 1.3 million comparisons, 99.9% of them yielded more than 1000 IBS 0 sites across the genome, with the highest number being 444,117. Out of the 156 comparisons with less than 1000 IBS 0 sites, 136 of the pairs were predicted by KING to be full siblings. While this relationship is impossible due to our population's non-overlapping generations, it suggests there is some underlying relatedness between these pairs of individuals or that KING is overestimating relatedness in our population. Overall, ~95% of all of

these comparisons obtained more IBS 0 sites then the highest amount generated by comparing all PO duos.

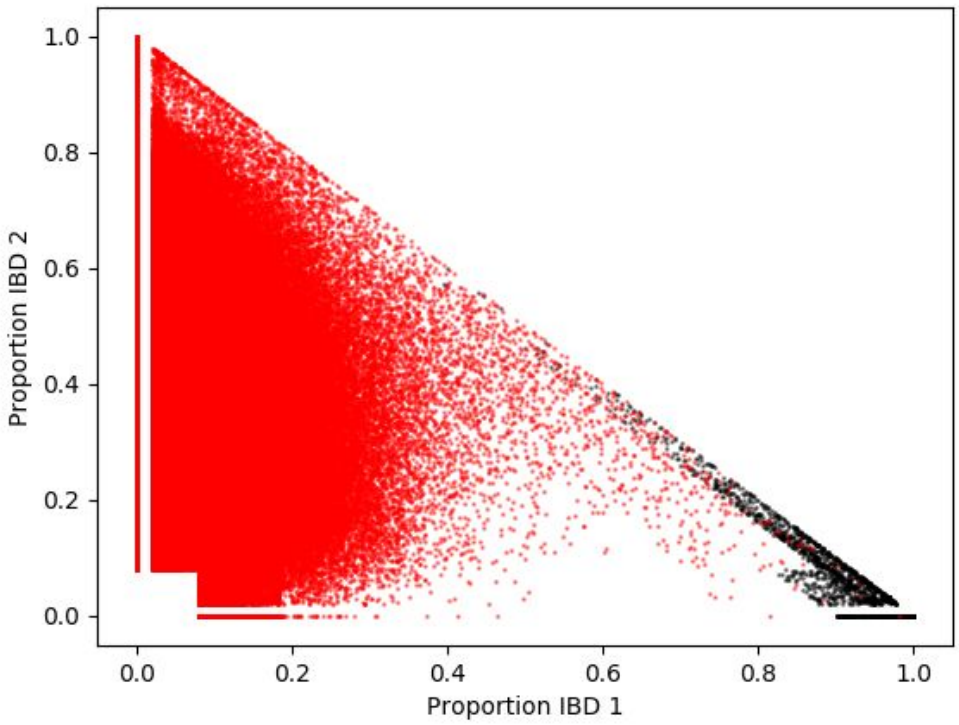

Figure 5: The proportion of the genome that is IBD 1 vs IBD 2 inferred by KING's ibdseg package. Black points represent the valid PO pairs utilized for the analysis. Red points are all of the comparisons between offspring of the valid PO duos and all other individuals in their parent's generation that were not also predicted to bet their parent.

In addition to the previous comparisons, I compared the IBD 1 and IBD 2 proportions between PO and non-PO pairs. The expected values of IBD 1 and IBD 2 for a true PO relationship are 1 and 0, respectively. Therefore the PO relationships predicted by KING should be within a range of these scores. The parental and non-parental pairs formed two distinct clusters when plotting IBD 1 vs IBD 2 (**Fig.**

5). However, the clusters are not discrete and have some overlaps. A small number of non-parental pairs displayed IBD levels expected of PO relationships. Additionally, a minor portion of the PO relationships had IBD 1 and IBD 2 values that were nearly equivalent. Equal portions of IBD 1 and 2 shared between two individuals does not represent the expected values of any of the common familial relationships presented in Table 1. This suggests that KING is overestimating identity by descent within our population, which is likely due our population's underlying structure. KING makes use of population allele frequencies to estimate likelihoods of a set of alleles being IBD 0,1, or 2. And since population allele frequencies in an admixed population do not necessarily represent allele frequencies within the respective founding populations, KING's model could be overestimating likelihoods of IBD and therefore relatedness.

Discussion

In this study, we utilized a robust library preparation method to successfully sequence nearly every member in the first four generations of our population in a cost-effective and timely manner. Our method involved producing unique dual indexed libraries for each individual in our population and then pooling 384 libraries at once for low coverage sequencing. On average, a sample within our population had a genome coverage slightly larger than 1x. Several recent studies have shown that this value is more than sufficient enough to perform meaningful analyses within low coverage data sets [Homburger et al., 2019; Navon et al., 2013; Pasaniuc et al., 2012].

Furthermore, the average coverage is in a range in which STITCH has been shown to perform imputation accurately. Using this imputation method we were able to generate a data set consisting of approximately 1.8 million confidently imputed SNPs and 3754 individuals. This informative and robust data set encapsulates genome wide information for the majority of sequenced flies. Regardless of a complete pedigree this data set can still be used to perform a variety of meaningful analyses such as ancestry and admixture inference.

When analyzing the rates of concordance between STITCH and bcftools genotype caller, we observed rates that suggest a lack of imputation accuracy. However, upon closer inspection the majority of discordant sites were the result of a homozygous call by bcftools and a heterozygous call by STITCH. This is likely due to the fact that our highest coverage samples are still sequenced to low depths and are naturally prone to genotype calling error. In particular, errors that result from the undersampling of both alleles at a diploid locus. It is less likely to sample every allele at a site under the constraints of low coverage sequencing, which would lead to an abundance of homozygous calls. Resequencing these individuals at higher coverage, 15x or greater [Song & Zhang, 2016], and then repeating the analysis would likely increase the rate of concordance. The only male included in the analysis had no instances of the previous discordance described above on the X chromosome. This is mostly due to the fact that since males only have one copy of the X chromosome, all male heterozygous calls on the X chromosome were masked prior to this analysis.

However, it is still interesting that ~85% of the discordance was due to heterozygous calls by bcftools at sites where STITCH imputed the male to be homozygous. This suggests that STITCH is imputing X chromosome genotypes accurately in low coverage males, whereas genotype calling can be error prone. Further support for this can be found when analyzing the unfiltered imputations on the X chromosome. On average ~97.9% of a male's imputations were homozygous, while females exhibited a much higher rate of imputation heterozygosity with an average of 68% of imputations being homozygous (**Fig. S1**).

In attempting to construct a population pedigree we used the imputed genotypes to complete pairwise relationship inference to identify PO relationships. This approach reduces the complexity of the pedigree by using the discrete nature of each generation within our population. Furthermore, the proportion of the genome that is shared by descent is unique to this type of relationship and can easily be validated by looking at allelic states. Using the software KING, we identified over one thousand confident and biologically valid PO relationships. However, we also identified several thousand PO duos that resulted in more than two parents per offspring, or offspring with parents of the same sex. Several thousand PO relationships that exist within the same generation were also ascertained. While we are able to construct some familial relationships, these erroneous parentage assignments make it impossible to confidently construct a whole population pedigree as of now.

KING's computational speed and ability to handle large multi sample data sets made it an attractive option for use in relationship inferences. However, KING's underlying algorithm utilizes population allele frequencies to calculate the likelihood of a set of alleles with the same state being shared by descent. Given our population's admixed structure, it is likely that allele frequencies within our population as a whole are not indicative of the different allelic frequencies that exist in the founding French and Ethiopian lineages. Due to this, KING may be overestimating the relatedness between individuals, thus making our relationship inference error prone [Sethuraman et al., 2018]. Recently, groups have developed software to estimate relatedness given a population mixed ancestry or population structure [Conomos et al., 2016; Moltke & Albrechtsen, 2014]. Softwares such as relateADMIX [Moltke & Albrechtsen, 2014] first identify the allelic frequency of the founding populations and use the ancestry inference of each individual to refine estimates of identity by descent. Unfortunately these software are not as robust and can lack the computational efficiency of softwares like PLINK or KING. Carefully pruning our data and reducing complexity may help to rectify these issues.

In conclusion, using our approach of low-coverage sequencing and genotype imputation we captured genome wide information for the vast majority of our sequenced population. We were able to use this information to identify a subset of familial relationships within our population. In order to improve the certainty of our results and continue to build the pedigree, we must explore other options that take our

population's mixed ancestry into account. Continuing to work towards constructing the complete pedigree, will increase the value of our current data set by adding extra information about the dynamic genetic history of our population. This could provide more power to traditional population genetic analyses.

Appendices

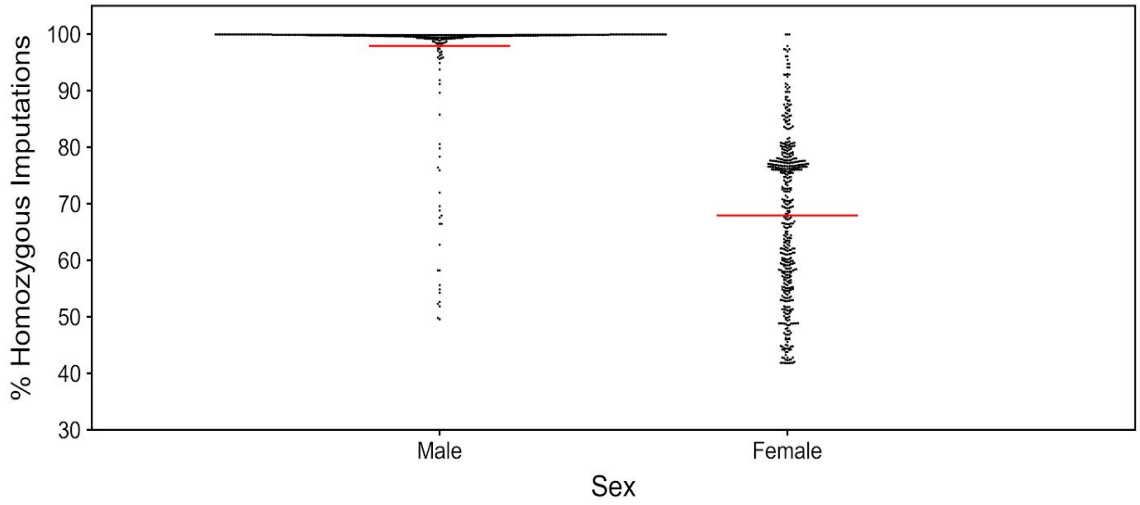

Figure S1: A swarm plot displaying the distribution between males and females of the percentage of homozygous imputations by STITCH for an individual along the X chromosome.. A random sample of 500 males and 500 females are represented in this figure, with each point representing one individual. The red lines indicate the mean percent of homozygous imputations within each random sample.

Bibliography

Browning, B. L., & Browning, S. R. (2013). Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics*, *194*(2), 459. https://doi.org/10.1534/genetics.113.150029

Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics*, *98*(1), 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020

Chen, N., Juric, I., Cosgrove, E. J., Bowman, R., Fitzpatrick, J. W., Schoech, S. J., Clark, A. G., & Coop, G. (2019). Allele frequency dynamics in a pedigreed natural population. *Proceedings of the National Academy of Sciences*, *116*(6), 2158–2164. https://doi.org/10.1073/pnas.1813852116

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free Estimation of Recent Genetic Relatedness. *American Journal of Human Genetics*, *98*(1), 127. https://doi.org/10.1016/j.ajhg.2015.11.022

Davies, R. W., Flint, J., Myers, S., & Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, *48*(8), 965–969. https://doi.org/10.1038/ng.3594

dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., Emmert, D. B., & Gelbart, W. M. (2015). FlyBase: Introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, *43*(D1), D690–D697. https://doi.org/10.1093/nar/gku1099

Epstein, M. P., Duren, W. L., & Boehnke, M. (2000). Improved Inference of Relationship for Pairs of Individuals. *American Journal of Human Genetics*, *67*(5), 1219. https://doi.org/10.1016/S0002-9297(07)62952-8

Hennig, B. P., Velten, L., Racke, I., Tu, C. S., Thoms, M., Rybin, V., Besir, H., Remans, K., & Steinmetz, L. M. (2018). Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. *G3: Genes|Genomes|Genetics*, *8*(1), 79. https://doi.org/10.1534/g3.117.300257

Heritable genetic alterations in a xeroderma pigmentosum group G/Cockayne syndrome pedigree. (1997). *Mutation Research/DNA Repair*, *385*(2), 107–114. https://doi.org/10.1016/S0921-8777(97)00031-1

Hollocher, H., Ting, C.-T., Pollack, F., & Wu, C.-I. (1997). Incipient Speciation by Sexual Isolation in *Drosophila Melanogaster*: Variation in Mating Preference and Correlation Between Sexes. *Evolution*, *51*(4), 1175–1181. https://doi.org/10.1111/j.1558-5646.1997.tb03965.x

Homburger, J. R., Neben, C. L., Mishne, G., Zhou, A. Y., Kathiresan, S., & Khera, A. V. (2019). Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Medicine*, *11*(1), 1–12. https://doi.org/10.1186/s13073-019-0682-2

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Kerr, S. M., Campbell, A., Murphy, L., Hayward, C., Jackson, C., Wain, L. V., Tobin, M. D., Dominiczak, A., Morris, A., Smith, B. H., & Porteous, D. J. (2013). Pedigree and genotyping quality analyses of over 10,000 DNA samples from the Generation Scotland: Scottish Family Health Study. *BMC Medical Genetics*, *14*(1), 1–7. https://doi.org/10.1186/1471-2350-14-38

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 1–13. https://doi.org/10.1186/s12859-014-0356-4

Lachance, J., & True, J. R. (2010). X-Autosome Incompatibilities in *Drosophila Melanogaster*: Tests of Haldane's Rule and Geographic Patterns Within Species. *Evolution*, *64*(10),3035–3046. https://doi.org/10.1111/j.1558-5646.2010.01028.x

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., Langley, C. H., & Pool, J. E. (2015). The *Drosophila* Genome Nexus: A Population Genomic Resource of 623 *Drosophila melanogaster* Genomes, Including 197 from a Single Ancestral Range Population. *Genetics*, *199*(4), 1229. https://doi.org/10.1534/genetics.115.174664

Lathrop, G. M., Hooper, A. B., Huntsman, J. W., & Ward, R. H. (1983). Evaluating pedigree data. I. The estimation of pedigree error in the presence of marker mistyping. *American Journal of Human Genetics*, *35*(2), 241.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987. https://doi.org/10.1093/bioinformatics/btr509

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. https://arxiv.org/abs/1303.3997v2

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078. https://doi.org/10.1093/bioinformatics/btp352

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834. https://doi.org/10.1002/gepi.20533

Lü, Q., Zhang, Y., Song, C., An, Z., Wei, S., Huang, J., Huang, L., Tang, L., & Tong, N. (2016). A novel SLC12A3 gene homozygous mutation of Gitelman syndrome in an Asian pedigree and literature review. *Journal of Endocrinological Investigation*, *39*(3), 333–340. https://doi.org/10.1007/s40618-015-0371-y

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867. https://doi.org/10.1093/bioinformatics/btq559

Moltke, I., & Albrechtsen, A. (2014). RelateAdmix: A software tool for estimating relatedness between admixed individuals. *Bioinformatics*, *30*(7), 1027–1028. https://doi.org/10.1093/bioinformatics/btt652

Navon, O., Sul, J. H., Han, B., Conde, L., Bracci, P. M., Riby, J., Skibola, C. F., Eskin, E., & Halperin, E. (2013). Rare Variant Association Testing Under Low-Coverage Sequencing. *Genetics*, *194*(3), 769–779. https://doi.org/10.1534/genetics.113.150169

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443. https://doi.org/10.1038/nrg2986

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., … Marchini, J. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics*, *10*(4). https://doi.org/10.1371/journal.pgen.1004234

Ouweland, J. M. W. van den, Lemkes, H. H. P. J., Ruitenbeek, W., Sandkuijl, L. A., Vijlder, M. F. de, Struyvenberg, P. a. A., Kamp, J. J. P. van de, & Maassen, J. A. (1992). Mutation in mitochondrial tRNA Leu(UUR) gene in a large pedigree with maternally transmitted type II diabetes mellitus and deafness. *Nature Genetics*, *1*(5), 368–371. https://doi.org/10.1038/ng0892-368

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., Sullivan, P. F., Bergen, S., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Purcell, S. M., Haas, D. W., Liang, L., … Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, *44*(6), 631–635. https://doi.org/10.1038/ng.2283

Picelli, S., Björklund, Å. K., Reinius, B., Sagasser, S., Winberg, G., & Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, *24*(12), 2033. https://doi.org/10.1101/gr.177881.114

Pool, J. E. (2015). The Mosaic Ancestry of the *Drosophila* Genetic Reference Panel and the *D. melanogaster* Reference Genome Reveals a Network of Epistatic Fitness Interactions. *Molecular Biology and Evolution*, *32*(12), 3236. https://doi.org/10.1093/molbev/msv194

Pool, J. E., Corbett-Detig, R. B., Sugino, R. P., Stevens, K. A., Cardeno, C. M., Crepeau, M. W., Duchen, P., Emerson, J. J., Saelao, P., Begun, D. J., & Langley, C. H. (2012). Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. *PLOS Genetics*, *8*(12), e1003080. https://doi.org/10.1371/journal.pgen.1003080

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, *81*(3), 559. https://doi.org/10.1086/519795

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, *22*(5), 939. https://doi.org/10.1101/gr.128124.111

Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F., & Neafsey, D. E. (2018). hmmIBD: Software to infer pairwise identity by descent between haploid genotypes. *Malaria Journal*, *17*(1), 1–4. https://doi.org/10.1186/s12936-018-2349-7

Sethuraman, A. (2018). Estimating Genetic Relatedness in Admixed Populations. *G3: Genes, Genomes, Genetics*, *8*(10), 3203–3220. https://doi.org/10.1534/g3.118.200485

Song, K., Li, L., & Zhang, G. (2016). Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Scientific Reports*, *6*(1), 1–7. https://doi.org/10.1038/srep35736

Ting, C.-T., Takahashi, A., & Wu, C.-I. (2001). Incipient speciation by sexual isolation in *Drosophila*: Concurrent evolution at multiple loci. *Proceedings of the National Academy of Sciences*, *98*(12), 6709–6713. https://doi.org/10.1073/pnas.121418898

Tore, S., Casula, S., Casu, G., Concas, M. P., Pistidda, P., Persico, I., Sassu, A., Maestrale, G. B., Mele, C., Caruso, M. R., Bonerba, B., Usai, P., Deiana, I., Thornton, T., Pirastu, M., & Forabosco, P. (2011). Application of a New Method for GWAS in a Related Case/Control Sample with Known Pedigree Structure: Identification of New Loci for Nephrolithiasis. *PLoS Genetics*, *7*(1). https://doi.org/10.1371/journal.pgen.1001281

Wu, M., Sun, L. V., Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J. C., McGraw, E. A., Martin, W., Esser, C., Ahmadinejad, N., Wiegand, C., Madupu, R., Beanan, M. J., Brinkac, L. M., Daugherty, S. C., Durkin, A. S., Kolonay, J. F., Nelson, W. C., Mohamoud, Y., … Eisen, J. A. (2004). Phylogenomics of the Reproductive Parasite Wolbachia pipientis wMel: A Streamlined Genome Overrun by Mobile Genetic Elements. *PLoS Biology*, *2*(3). https://doi.org/10.1371/journal.pbio.0020069

Yukilevich, R., & True, J. R. (2008). Incipient Sexual Isolation Among Cosmopolitan *Drosophila Melanogaster* Populations. *Evolution*, *62*(8), 2112–2121. https://doi.org/10.1111/j.1558-5646.2008.00427.x

Zan, Y., Payen, T., Lillie, M., Honaker, C. F., Siegel, P. B., & Carlborg, Ö. (2019). Genotyping by low-coverage whole-genome sequencing in intercross pedigrees from outbred founders: A cost-efficient approach. *Genetics Selection Evolution*, *51*(1), 1–11. https://doi.org/10.1186/s12711-019-0487-1